WaveFontStyler: Font Style Transfer Based on Sound

Kota Izumi[0000−0001−7374−9205] and Keiji Yanai[0000−0002−0431−183X]

The University of Electro-Communications, Tokyo, Japan
{izumi-k,yanai}@mm.cs.uec.ac.jp

Abstract. This study proposes a method for transferring a style aligned with the input sound to character images. By combining ImageBind, a model that embeds sound and image features into the same space, with a Bezier curve-based text representation, the style corresponding to the input sound is effectively transferred to the character images. Experiments demonstrated the successful generation of character images that reflect the sound-aligned style. Additional results with input sounds can also be viewed and heard at the following link: https://squ602.github.io/WaveFontStyler/

Keywords: font style transfer · neural style transfer · differentiable renderer · large-scale multimodal model.

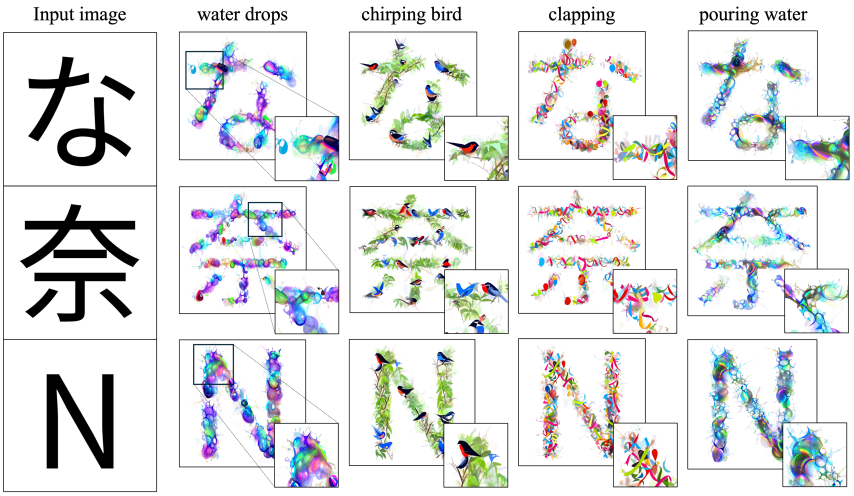

Fig. 1: Results of the stylistic transformation of the characters using this method. In addition, enlarged views of specific parts of each character are also presented.

1 Introduction

Artistic fonts are widely used in advertisements, posters, and web pages. However, designing artistic text varies significantly from character to character and requires significant effort, as different types of characters, such as kanji, katakana, and others, need to be designed separately.

Recent advancements in deep learning have simplified the process of creating artistic fonts by specifying conditions through text. Nevertheless, there are instances where it is desirable to express subtle nuances that text alone cannot capture. Utilizing sounds to specify such detailed nuances could help achieve a style more closely aligned with our vision. As a first step, this study explores the potential of sound input in expressing various styles.

Recently, ImageBind [4] has been proposed as a method to embed features of six modalities, including sound, image, and text, into a common embedding space. By obtaining embeddings using the encoders trained with this approach, the cosine distance between sounds, images, and texts within the ImageBind embedding space can be measured.

In this study, we propose a method that uses ImageBind to transfer a style aligned with the input sound to the text image. Text is represented using Bezier curves, with the parameters of each curve being directly updated to achieve a style closer to the specified one. Experiments demonstrate the ability to generate character images that reflect the sound-aligned style.

2 Related Works

2.1 ImageBind

Recently, ImageBind [4] has been proposed as a method to embed multiple modalities into a common space. This method learns common embeddings across six different modalities: images, text, sound, depth, thermal, and IMU data. Instead of using paired data for all modalities, it trains using paired data between images and other modalities (e.g., image-text, image-sound). ImageBind performs on par with or even outperforms models trained with supervised data in zero-shot sound classification and retrieval tasks. Furthermore, it can generate images from sound when integrated with a large-scale image generation model.

In this study, we measure the cosine distance between images and sounds within the ImageBind space to calculate their similarity.

2.2 Generation and transformation using differentiable renderers

In recent years, research combining differentiable renderers [8] with large-scale foundation models has been actively pursued. Differentiable renderers enable the backpropagation of loss, calculated using raster images, to the parameters of primitives that constitute vector images by making the rendering process from vector images to raster images differentiable. In CLIPDraw [2], the parameters of

Bézier curves are optimized using CLIP [10] to generate images solely from text. CLIPFont [12] takes vector-format character images as input and transfers styles that match the given prompt to the characters by optimizing the parameters of the Bézier curves representing the characters, using CLIP. Furthermore, Word-as-Image [5] deforms the shape of characters to reflect the meaning of a prompt by optimizing the contours of characters represented by Bézier curves, utilizing a pre-trained diffusion model [11]. Additionally, in the method proposed by Izumi *et al.* [6], they achieved style transfer for characters using prompts by optimizing Bézier curves with CLIP while imposing constraints on the shapes.

Based on these studies, our research combines a differentiable renderer with ImageBind to optimize the parameters of Bézier curves, enabling the generation of fonts where styles specified by sound are transferred while preserving the character shapes.

3 Method

An overview of the proposed method is shown in Figure 2. In this method, images are represented not as pixel-based raster images but as a set of closed Bézier curves, where each curve's position and color are controlled by parameters. These Bézier curve parameters are directly optimized using gradient descent. Additionally, a differentiable renderer [8] is employed to propagate the loss, calculated from the rasterized image, back to the Bézier curve parameters via gradient descent. For loss computation, we use a pre-trained ImageBind [4] model to calculate the cosine distance between the rendered image and the input sound in the ImageBind space. Furthermore, to preserve the shape of the characters, we incorporate the Distance Transform Loss, introduced by Atarsaikhan *et al.* [1] for font style transfer.

3.1 Drawing Representation

The vector images are composed of a set of closed Bézier curve paths, based on the method by Li *et al.* [8]. Each path consists of multiple segments, with parameters for the positions of control points, fill color, and opacity. These parameters are optimized using gradient descent. The initial positions of the control points are determined in the same manner as in the method proposed by Izumi *et al.* [6] as shown in Figure 3(b), the coordinates of black areas in the character image are randomly sampled according to the number of strokes to set the coordinates of the starting control points for each path. This approach is employed to prevent Bézier curves from being drawn over irrelevant background areas, which can occur if random positions in the image are used as initial points.

After determining the starting control points, the remaining control points are randomly positioned around the initial control points.

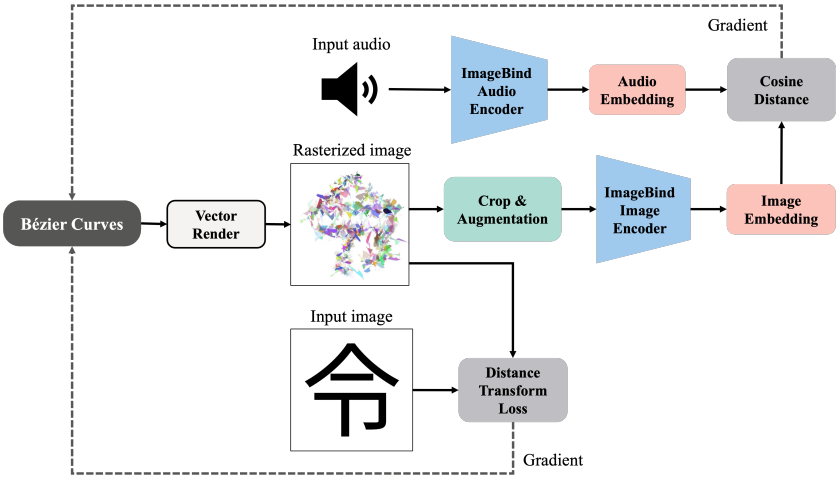

Fig. 2: Overview of the proposed method.

(a) (b) (c)

Fig. 3: (a) Input image (b) The initial positions of Bézier curves. Control points are distributed near these points. (c) The distance transform map. The further away from the text area, the larger the value.

3.2 Loss Functions

The loss function includes a Distance Transform Loss to preserve the shape of the characters and a Similarity Loss to transfer the style specified by the sound to the characters.

Distance Transform Loss
To transfer styles to fonts, it is necessary to preserve the shape of the characters. To achieve this, our method incorporates the Distance Transform Loss, which was introduced by Atarsaikhan *et al.* [1] for style transfer to logo parts. To compute the Distance Transform Loss, we first perform a distance transform on the input image to create a distance-transformed image. Then, the distance-transformed images are multiplied with both the input and output images, and the mean squared error between these products is calculated. As shown in Figure

3(c), the distance transform sets the pixel values of the character region to 0, with higher values assigned to pixels farther from the character's outline. The Distance Transform Loss can be defined by the following equation.

$$L_{\text{distance}} = \frac{1}{2} \left(I_c \circ I_d - I_{draw} \circ I_d \right)^2 \tag{1}$$

Here, I_c, I_d, I_{draw} represent the input image, the distance-transformed image, and the rendered image, respectively. The operator $\circ$ denotes the element-wise multiplication of matrices of the same size.

Similarity Loss

To transfer the style specified by sound to the characters, we calculate the cosine distance in the ImageBind embedding space between the rendered image and the input sound. Specifically, following CLIPStyler [3], we randomly extract N patches from the image rendered using a differentiable renderer. In addition, similar to CLIPDraw [2], we apply a perspective transformation to the patches. Then, using the Image Encoder trained with the ImageBind approach, each patch is encoded into a 1024-dimensional vector. The input sound is also encoded into a 1024-dimensional vector using a similarly trained encoder. The cosine distance between these vectors is then calculated. Finally, the average cosine distance across all patches is computed. Therefore, the Similarity Loss can be defined as follows.

$$l^i_{\text{patch}} = 1 - \frac{E_I(\text{aug}(\hat{I}^i_{draw})) \cdot E_A(A)}{|E_I(\text{aug}(\hat{I}^i_{draw}))\| E_A(A)|}$$

$$L_{\text{sim}} = \frac{1}{N} \sum_i^N l^i_{\text{patch}} \tag{2}$$

Here, $\hat{I}^i_{draw}$ represents the i-th patch extracted from the rendered image, and aug denotes the perspective transformation. E_I and E_A represent the Image Encoder and Audio Encoder, respectively, both of which are pre-trained using the ImageBind approach.

Total Loss

Using the two loss functions described above, the final Total Loss is defined as follows.

$$L_{\text{total}} = \lambda_{\text{dist}} L_{\text{dist}} + \lambda_{\text{sim}} L_{\text{sim}} \tag{3}$$

3.3 Reinitialization and deletion of paths

During optimization, opacity may decrease or the path area may shrink, resulting in some paths contributing minimally to the rendering. In the case of characters, parts of the characters may become faint. To address this issue, VectorFusion [7]

periodically reinitializes paths with opacity or area below a certain threshold. Our method also adopts this approach.

Additionally, for characters, there is a phenomenon where small, faint paths unrelated to the intended style extend beyond the character boundaries. To mitigate this, our method not only reinitializes paths but also removes paths with opacity or area below a threshold at the end of the optimization process.

4 Experiments

4.1 Experimental Setting

The resolution of the input images was set to 512×512. The weights λ_{dist} and λ_{sim} were set to 1.0 and 0.5, respectively. These weights were chosen to best reflect the style while maintaining the readability of the characters. The optimization function used was Adam, with learning rates set to different values for the position of the control points and the color (including opacity): 1.0 and 1.0×10^{-2}, respectively. The number of iterations was set to 100, and the size of the cropped images was set to 160×160. Additionally, for the calculation of the loss function, 24 patches were extracted from the rendered image.

Each character was represented using 512 paths. The generation time was approximately 2 minutes per image on an NVIDIA RTX A6000. The input sounds were taken from the environmental sound dataset ESC-50 [9].

4.2 Qualitative evaluations

The results of our method are shown in Figure 1. It can be seen that the character images have been transformed to match the style of the input sound. For example, when "water drop" sound is used as input, the style appears to have a texture similar to colored water droplets falling onto paper. When "pour water" sound is used, a texture resembling light reflections appears, indicating a different style from the one produced with the water-dropping sound. Additionally, when the "chirping bird" sound is used as input, the image reflects birds perching in trees. In the case of "clapping", elements resembling party poppers are present, possibly evoking a celebratory image.

5 Conclusion

In this study, we proposed a method for transferring styles corresponding to input sounds to character images. By combining ImageBind with a Bézier curve-based representation of characters, our experiments demonstrated that the style aligned with the input sound could be effectively transferred to the character images.

Acknowledgments This work was supported by JSPS KAKENHI Grant Numbers, 22H00540 and 22H00548.

References

1. Atarsaikhan, G., Iwana, B.K., Uchida, S.: Contained neural style transfer for decorated logo generation. In 13th IAPR International Workshop on Document Analysis Systems (2018)
2. Frans, K., Soros, L., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843 (2021)
3. Gihyun, K., Chul, Y.J.: Clipstyler: Image style transfer with a single text condition. Proc.of IEEE Computer Vision and Pattern Recognition pp. 18062–18071 (2022)
4. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. Proc.of IEEE Computer Vision and Pattern Recognition (2023)
5. Iluz, S., Vinker, Y., Hertz, A., Berio, D., Cohen-Or, D., Shamir, A.: Word-as-image for semantic typography. ACM Trans. Graph. **42**(4) (jul 2023)
6. Izumi, K., Yanai, K.: Zero-shot font style transfer with a differentiable renderer. Proc.of the 4th ACM International Conference on Multimedia in Asia (32), 1–5 (2022)
7. Jain, A., Xie, A., Abbeel, P.: Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. Proc.of IEEE Computer Vision and Pattern Recognition (2023)
8. Li, T.M., Lukáč, M., Gharbi, M., Ragan-Kelley, J.: Differentiable vector graphics rasterization for editing and learning. ACM Transactions on Graphics (Proc. SIGGRAPH Asia) **39**(6), 193:1–193:15 (2020)
9. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. Proceedings of the 23rd Annual ACM Conference on Multimedia pp. 1015–1018 (2015)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Pamela Mishkin, J.C., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. Proc.of the 38th International Conference on Machine Learning **139**, 8748–8763 (2021)
11. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. Proc.of IEEE Computer Vision and Pattern Recognition (2022)
12. Song, Y., Zhang, Y.: CLIPFont: Text guided vector wordart generation. 33rd British Machine Vision Conference (2022)